# Une base de données sur les troncations involontaires de mots en français parlé.

Berthille Pallaud

# UNE BASE DE DONNÉES SUR LES TRONCATIONS INVOLONTAIRES DE MOTS EN FRANÇAIS PARLÉ

Berthille Pallaud

## 1. Introduction

C'est à partir d'études sur la production et la perception « d'erreurs de langage » (Pallaud, 2001, 2002) dans des corpus de français parlé que s'est imposée la nécessité d'une clarification de ces phénomènes si fréquents à l'oral et de la constitution d'une base de données propre aux troncations de mots. Cette base de données a permis une analyse approfondie des amorces de mots identifiées dans des énoncés oraux transcrits et a également permis de préciser les apports de la notion de disfluence[1] et les effets du développement, dans ce secteur, des nouvelles technologies, en particulier lors des dix dernières années.

## 2. Du lapsus aux disfluences

Le lapsus, que le sens commun (du moins si on en juge par ce qu'en disent les média, par exemple) relie sans hésiter au domaine de la psychanalyse, est un terme présent dans tous les dictionnaires de psychanalyse mais absent des lexiques ou dictionnaires de linguistique actuels. Pourtant, comme on va le voir, certains linguistes se sont intéressés aux lapsus, en particulier depuis les années 80 aux États-Unis, et quelque dix ans plus tard en France. Ce renouveau d'intérêt tient aux évolutions théoriques dans le domaine de la linguistique (surtout en phonétique) et celui de la psycholinguistique, majoritairement sous l'impulsion du cognitivisme. En revanche, la conception des études et spécifiquement la constitution de la base de données permettant les analyses, sont restées conformes à la tradition : elles sont fondées sur un recueil patient et souvent sporadique (dit « à la volée ») des phénomènes recherchés.

PALLAUD, Berthille (2006), Une base de données sur les troncations involontaires de mots en français parlé, *Travaux Interdisciplinaires du Laboratoire Parole et Langage,* vol. 25, p. 173-184.

1. Le terme de disfluence, employé pour décrire les trébuchements de la parole dans les énoncés de français parlé *standard* est à l'origine un anglicisme.

C'est ainsi que le philologue Meringer et le psychiatre Mayer avaient procédé en leur temps (1895, 1908, 1923) à la collecte des matériaux de leur étude. Ils sont considérés comme ayant été les premiers à avoir véritablement étudié les lapsus et constitué à la fin du XIX[e] siècle un corpus de 8800 lapsus, dont la moitié sont des *lapsus linguae* (les autres sont des *lapsus calami* et *lectionis*)[2]. Il n'est pas une étude traitant des lapsus ou ratés de langage qui ne fasse référence aux travaux de ces auteurs allemands.

Les chercheurs qui ont étudié les lapsus en langue française, dans les années 90, ont constitué des recueils limités au cadre strict de la production langagière orale. Ils ont utilisé une méthode qui s'apparente au recueil à la volée[3] puisqu'ils s'appuient sur la manipulation de longs corpus (Rossi et Peter-Defare, 1998) et non sur l'extraction systématique de phénomènes à partir de transcriptions de corpus.

Au Laboratoire Parole et Langage, une série de recueils et d'études a abouti à plusieurs publications dont les principales sont celles de Rossi et Peter-Defare (1995, 1998) qui portent sur environ 4000 exemples d'erreurs de langage. Le travail d'E. Peter-Defare (1993) est la première étude systématique et linguistique dans ce domaine en langue française. Cette base de données est accessible sur le site du LPL[4].

Arnaud (1997) a publié également une première analyse sur son propre recueil de lapsus qu'il qualifie de « naturels » (pour les distinguer des lapsus déclenchés en laboratoire, ou suscités artificiellement). La notation de ces 2400 exemples a été faite sur le mode « systématique » *(sic)* (« tout lapsus entendu a été noté », p. 308) et cela dans les secondes qui suivaient leur production (« afin d'éviter la rapide dégénérescence de la trace mémorielle », p. 308).

Fénoglio (1997), dans une approche quasi clinique pour aborder l'énonciation en discours oral, s'est appuyée sur un corpus de 25 lapsus relevés dans des entretiens enregistrés. Le lapsus apparaît comme une donnée d'articulation entre discours et parole.

Les travaux de Pillon (1998), à l'aide de lapsus déclenchés artificiellement, montrent l'implication de la morphologie dérivationnelle des mots dans les processus de production de la parole.

---

2. *Lapsus linguae* : lapsus oral ; *lapsus calami* : lapsus écrit ; *lapsus lectionis* : lapsus en lecture.

3. « Il s'agit de lapsus relevés soit dans les conversations de la vie courante, soit dans des dialogues et tables rondes diffusés sur les médias entre 1992 et 1996. Nous n'avons retenu que les lapsus dont nous étions sûrs et pour lesquels nous pouvions obtenir un contexte suffisant où pouvait être identifiée l'origine éventuelle. » (Rossi & Peter-Defare, 1998, p.17).

4. Liens : Outils/Bases de Données/Lapsus en français : <http://www.lpl.univ-aix.fr/index.php?id=240>.
Une quarantaine de lapsus sont accessibles. Les collègues qui souhaiteraient avoir accès à la totalité de la base pour des besoins de recherche sont invités à contacter Mario Rossi.

Ainsi, mis à part les travaux sur les lapsus déclenchés artificiellement, les études sur les lapsus ont en commun d'avoir été construites sur un ensemble d'exemples recueillis plus ou moins à la volée avec, la méthode l'implique, une définition *a priori* de ce que signifie un lapsus : à savoir une erreur involontaire de langage, se manifestant (sous l'effet d'un processus inconscient) par la production d'un un mot pour un autre, incongru ou déformé.

- Un mot pour un autre :
  *oh mon écharpe + je l'ai prêtée à Gisèle parce qu'elle avait froid et* **je** *l'ai gardée euh* **elle** *l'a gardée*
  (corpus Pallaud 1997, (56)

- Un mot incongru :
  L1- *tu as entendu le Sud du Portugal est relié par le plus grand* **port** *d'Europe*
  L2- *tu veux parler d'un* **pont** *je suppose*
  L1- *oui bien sûr* (corpus Pallaud 1997, 125)

- Un mot déformé :
  *ma mère a dit surtout cache ce livre que Poupette ne le lise pas eh bien ce livre je l'ai* **[li]** *caché dans le lit*
  (corpus Pallaud 1997, 124)

Cette méthode prévoit de constituer des recueils *chemin faisant* au gré des conversations ou émissions de radio écoutées par les collecteurs[5]. Ne reposant pas sur des corpus oraux transcrits, elle ne prétend pas être exhaustive. Par ailleurs, l'identification même de ces phénomènes relève tantôt d'une analyse sémantique (l'incongruité d'un terme) tantôt de la détection des transgressions de règles morphologiques (un mot déformé).

Dans une étude entreprise sur les lapsus (Pallaud, 1999) une comparaison a pu être conduite sur deux types de recueils de disfluences : l'un à la volée, l'autre rassemblant les achoppements extraits (de façon systématique et exhaustive) d'un corpus enregistré (corpus Bertuzzi, 1997[6] et transcrit selon les conventions du Groupe Aixois de Recherche en Syntaxe (GARS)[7]. Dans les deux cas, les énoncés sont des propos tenus dans le cadre d'entretiens ou de conversations. L'étude du corpus Bertuzzi (1997) s'est révélée essentielle car la possibilité de lire et de relire la transcription des paroles prononcées a permis d'éviter de construire une analyse sur des faits déjà définis (par

5. Le relevé de lapsus est plus facile à faire lors d'écoutes d'émissions radio. Il n'est, en effet, pas facile d'adopter une attitude d'observateur tout en participant à une conversation avec d'autres personnes (Gadet, 2000, 2003).
6. Il s'agit du corpus Bertuzzi 97 qui est la transcription intégrale de trois enregistrements d'une personne évoquant les conditions d'exercice de son métier de fleuriste à Paris. Cette étude a été faite dans le cadre d'une soutenance de maîtrise au département de linguistique française à l'université d'Aix-Marseille I.
7. Ce qu'Arnaud (1997) nomme recueils extensif (à la volée) et intensif (sur corpus enregistré).

exemple, du type « un lapsus est un mot pour un autre »[8]). La lecture de l'énoncé permet de constater des phénomènes qui sont bien sûr perceptibles pour une écoute attentive et avertie mais souvent impossibles à noter à la volée. L'étude sur corpus présente l'avantage de permettre surtout un relevé systématique de faits qui autrement paraissent rares (parce que discrets et rapides). Il est évident que le recueil à la volée n'est pas systématique (puisqu'il est impossible de noter tout achoppement entendu), mais, de plus, il est apparu qu'il a évolué sous l'influence des constats faits lors de la comparaison des recueils. En d'autres termes, si au départ cette activité de recueil semblait simple, l'analyse du corpus enregistré a révélé des phénomènes insoupçonnés. Cela ne manqua pas de modifier le recueil à la volée ; ces phénomènes, passés inaperçus dans la vie quotidienne, devenaient faciles à relever. S'il y a une certitude méthodologique, c'est évidemment qu'un tel recueil de faits linguistiques ne peut être ni constant ni systématique (c'est bien ce qu'a souligné Gadet dans ses réflexions « *Derrière les problèmes méthodologiques du recueil des données* », 2003).

En particulier, bien que la fréquence des amorces de mots puisse beaucoup varier d'un corpus à l'autre, cette étude (Pallaud, 1999) a montré le biais qui peut, vraisemblablement, être introduit par un recueil à la volée et la nécessité qu'il peut y avoir à bénéficier de corpus enregistrés si on veut étudier la place de ces phénomènes dans l'élaboration des énoncés oraux. En effet, 59% des erreurs de langage dans le corpus enregistré sont des amorces de mots alors que dans ces types de recueils, le pourcentage tombe à 10,8%. En conséquence, les erreurs du type « un mot pour un autre » sont, sans doute, très surestimées dans les recueils à la volée, alors que les hésitations se traduisant par des amorces (amorces inachevées, corrigées et complétées) se trouvent sous-estimées.

## 3. Du recueil à la volée à l'extraction des phénomènes linguistiques

Les nouvelles technologies, en facilitant les enregistrements et le stockage des données sonores, ont permis le développement d'une culture des ressources langagières[9]. Il n'a plus été question de recueil à la volée mais d'extraction automatique ou non et de descriptions systématiques dans des corpus souvent volumineux (évalués en millions de mots). Le terme de *lapsus* disparaît alors au profit de ce qui a été nommé disfluences c'est-à-dire toutes les marques de trébuchement dans les énoncés. En renonçant à une définition *a priori* du phénomène lapsus, on peut réinsérer ce type de

8. La langue fourche de bien des façons et les achoppements ne se limitent pas au glissement d'un mot vers un autre. De plus, à l'oral, la notion de mot (sans parler de la notion de la phrase) soulève également des problèmes d'identification.

9. La Délégation Générale à la Langue Française, par exemple, a financé, dans les années 90, la constitution d'un Corpus de Référence de Français Parlé sur toutes les régions de France (équipes aixoises dirigées par Claire Blanche-Benveniste puis Jean Véronis).

phénomènes dans le champ des spécificités de l'oral. Ce faisant, on en revient à l'origine du terme latin *lapsus* qui, tout comme celui qu'utilisent les anglo-saxons, *slip* (*of the tongue*) évoque le trébuchement, l'achoppement.
C'est qu'en effet, les études sur les corpus oraux montrent combien les vers de Boileau :

> *Ce qui se conçoit bien s'énonce clairement.*
> *Et les mots pour le dire vous viennent aisément.*

ne décrivent pas la situation d'énonciation. Le locuteur construit son énoncé au travers de différents procédés qui montrent à l'évidence que l'écoulement de la parole n'est pas fluide.
Si la fluidité d'un énoncé oral se mesure à la régularité rythmique dans sa production, il est clair que les énoncés oraux ne sont pas fluides mais disfluents (Shriberg, 1999 ; Pasdeloup, 1992). Tout locuteur produit de la parole avec une certaine variabilité dans le débit des mots, des pauses (silencieuses ou non, Duez, 2001), des allongements d'éléments linguistiques, *etc.* De façon générale, ces phénomènes, peu perçus par les locuteurs en présence (y compris par celui qui parle), déclenchent très rarement des commentaires. Certaines de ces disfluences dans la parole, au demeurant très fréquentes, se caractérisent par une interruption (notée IP) dans l'énoncé, que ce soit au niveau morphologique (l'amorce de mot) ou à la frontière de mots (ce qui est le cas du phénomène de bribe suivi ou non de répétition de mots ; Schriberg, 1999 ; Henry, 2002) :

(1) Amorce de mot : *c'est vrai que c'est pas* **b-***(IP)* **beau** *d'associer les deux* (Arborign, 5, 14)
(2) Bribe suivie de répétition : *ils auront* **leur propre** *(IP)* **leur propre** *langage* (Laurent, 1, 2)

Le résultat de la transcription d'un énoncé oral diffère en bien des points d'un énoncé écrit, du moins de la forme achevée d'un énoncé écrit[10]. Les locuteurs qui se sont prêtés au jeu de l'enregistrement et qui ont eu accès par la suite au texte transcrit ont été au mieux surpris mais le plus souvent choqués par « la façon dont ils ont parlé ».
Les avatars spécifiques à l'énoncé oral (donc, absents, sauf exception, des textes écrits) et repérables dans une transcription se limitant au texte du discours sont de divers ordres :

- les reprises ou faux départs
  *J'avais beaucoup de mal* **à à à** *gérer* **ce ce** *genre de difficulté(s)* (corpus C5aBelfo, 3,1)
- les pauses remplies ou silencieuses
  *C'est comme toute* **euh** *tout ce qu'on a envie de faire découvrir* (corpus C5bBelfo, 16,1)
- les incises autonymiques ou interjections

10. Les études sur les manuscrits ou les brouillons de textes publiés témoignent du patient travail d'écriture et de réécriture fourni par l'écrivain *(cf.* Lebrave).

**Ouais disons** *que ça a quand-même beaucoup beaucoup* **enfin** *pas ma- em-* **comment dire** *ça a énormément évolué* (corpus C41Cstra, 41,3)

- les interruptions de mots ou de syntagmes
  *C'est quand même tout à fait intéressant à* **ra-** *rappeler en passant* (corpus C41Astg, 3,1)
  *Comment ils appellent ça* **le liquide donc c'est mh le comptage du + le ni-** *comment dire ça le léchage euh* (corpus C41Cstg, 20,5)
- les ruptures syntaxiques
  **Là c'était c'était c'ét-** *il se trouvait que euh la fondation /était, est/ en Allemagne* (corpus Pariscen, 9,1)
- les remplacements de mots ou les néologismes souvent nommés *lapsus*
  *Si j'ai tant attendu pour* **attendre** *ce livre pour* **écrire** *ce livre* (Pallaud, 1997, 142)
  *j'ai l'impression que je vais* **apprir** *que je vais apprendre à vivre avec toi* (Pallaud, 1997, 28)

Comme l'ont souligné Clark et Wasow (1998), ces phénomènes, qui témoignent d'une difficulté dans la fluidité verbale et que les linguistes ont appelé disfluences (par opposition aux phénomènes dits pathologiques de disfluence)[11] ont en commun la même structure de fonctionnement : une interruption se produit dans la fluidité de la parole que le locuteur va dépasser de diverses manières. Cette interruption est marquée en son point de rupture par des phénomènes phonétiques (travaux de Schriberg, 1999, par exemple).

Les transcriptions qui adjoignent au texte la description de différents éléments prosodiques et phonétiques (comme l'intonation, l'accent, la durée des syllabes, par exemple) montrent que les paroles du locuteur diffèrent en certains points de ce qui est décrit pour un locuteur idéal dit « standard ». Ces marques acoustiques accompagnent en particulier, on l'a vu, les points d'interruptions des énoncés. Il reste qu'en dehors de ces passages, il est possible de trouver des disfluences intonatives ou phonétiques qui introduisent des changements de sens pour l'énoncé comme dans l'exemple suivant que l'on transcrirait et ponctuerait alors ainsi selon l'intonation adoptée :

*Ne reviens pas trop tard*
*Ne reviens pas ; trop tard.*

# 4. La base de données sur les troncations de mots

## 4.1. Les recueils de corpus

La base de données ne concerne que les phénomènes de troncation de mots. Ils ont été identifiés dans une partie du « Corpus de référence de français parlé ». La constitution de ce recueil de production langagière répondait à une requête de la Délégation à la Langue Française (ministère de

11. Par définition les *disfluences* sont la marque d'un dysfonctionnement dans la parole alors que les *disfluences* ne sont pas liées à une pathologie.

la Culture), qui l'a totalement financée. La réalisation de ce projet avait été confiée en 1998 à l'URA 6060 du CNRS (sous la responsabilité de Claire Blanche-Benveniste), et à partir de 2000, le projet a été pris en charge par la Jeune équipe DELIC (DEscription Linguistique Informatisée sur Corpus), dirigée par Jean Véronis. L'objectif de ce corpus était de mettre à la disposition de la communauté de linguistes des témoignages de la langue française parlée aujourd'hui, dans les principales villes de l'hexagone. Des enregistrements (132) de locuteurs choisis selon des critères précis (âge, niveau d'éducation, *etc.)*, dans une quarantaine de villes de France, représentant des types de parole variés (parole publique, professionnelle ou privée) sont transcrits selon les conventions du GARS. La totalité est estimée à environ 50 heures de parole, soit environ 400000 mots[12]. La transcription et la segmentation ont été réalisées majoritairement à l'aide du logiciel TRANSCRIBER 1.4.2.[13], programme développé par Claude Barras spécialement pour la transcription et l'alignement de corpus oraux.

Depuis les années 70, l'équipe du GARS d'Aix a développé pour ses études une méthode de transcription des corpus de français parlé contemporain (Blanche-Benveniste et Jeanjean, 1987). Les deux conventions principales de cette méthode, à savoir une transcription orthographique et l'absence de ponctuation, ont été adoptées par la plupart des concepteurs de méthodes actuelles pour la transcription et l'édition de grand corpus (Blanche-Benveniste, 1997). Le souci d'établir un document linguistique a prévalu et avec lui, le refus d'apurer le texte de l'oral de ses soi-disant scories, même si la lecture de ces documents peut sembler au premier abord incommode pour un lecteur non averti :

> *C'est l'accumulation des procédés caractéristiques des « avant-textes » de l'oral qui rend la lecture des transcriptions si incommode : retouches de toutes sortes, « hésitations », « énoncés inachevés », « ruptures de construction », pauses fréquentes,* etc. (Blanche-Benveniste et Jeanjean, 1987, p. 162).

Indications prosodiques et transcriptions phonologiques ne sont employées que dans certains passages (par exemple pour les énoncés ne pouvant recevoir une interprétation). Encore ne sont-elles pas portées dans le texte mais renvoyées en note (à l'exception des transcriptions d'enregistrement d'aphasiques à propos desquelles cette difficulté d'interprétation est si fréquente parfois que la lecture en serait trop incommode).

12. Pour le français parlé, il offre une base de comparaison avec les corpus de français parlé hors hexagone, le corpus Valibel en Belgique (dix fois plus important) et le corpus d'Ottawa-Hull au Canada ; il représente l'amorce de la réalisation d'un corpus comparable à celui du *British National Corpus,* pour l'anglais parlé.

13. Logiciel téléchargeable gratuitement sur le site : <http://www.etca.fr/CTA/gip/Projets/Transcriber/>

Selon les conventions du GARS, qui prévoient une transcription orthographique des énoncés oraux, les amorces de mots sont notées par un trait d'union collé au fragment du mot et donc identifiable automatiquement de façon univoque.
Tous ces enregistrements, sauf deux, ont été conduits en privé, selon la méthode de l'interview la moins directive possible, et ne rassemblaient que deux locuteurs adultes (l'interviewer et l'interviewé). Les deux corpus faits en public étaient des exposés improvisés devant un groupe de 40 personnes.

## 4.2. Les troncations involontaires de mots

A l'heure actuelle, les troncations de mots ont été extraites d'un sous-ensemble de 20 corpus recueillis, à l'exception de l'un d'entre eux, dans le cadre de l'enquête précitée. Il comporte 105000 mots. Ces corpus ne sont pas étiquetés. Afin de pouvoir analyser les énoncés du seul locuteur interviewé (et non ceux du locuteur interviewer), ils ont été extraits des 20 corpus-sources pour constituer un deuxième sous-ensemble de 20 corpus interviewés. Chaque corpus correspond alors à un seul locuteur. C'est dans ce deuxième sous-ensemble qu'ont été recherchées les amorces de mots, à l'aide du logiciel CONTEXTES (établi par J. Véronis). Ce logiciel a permis de recenser les amorces de mots présentes, entourées d'un contexte antérieur et postérieur de 10 à 30 mots chacune. Les énoncés contenant ces fragments de mots sont au nombre de 441.
Si l'on se fonde sur un débit moyen de 200 mots/mn, le nombre total de mots étant de 105000 mots, la durée totale de l'ensemble de ces corpus est de 7h51 mn. La longueur moyenne des corpus est de 3080 mots, soit une durée moyenne de 16 mn. ; les valeurs extrêmes de ces durées sont de 1307 et 4931 mots. On retrouve, dans ce sous-ensemble de corpus, la fréquence moyenne d'apparition des amorces de mots dans un énoncé (1 amorce/57s. ; variation individuelle de 1/23 s. à 1/8 mn.).
Les 441 énoncés extraits de ces corpus interviewés sont rassemblés dans un tableur EXCEL. L'étiquetage (non automatique) est fait, dans un deuxième temps : les 441 énoncés comportant les troncations de mots sont disposés dans une même colonne (événements). Chaque ligne ne contient qu'un exemple de troncation resitué dans son contexte. Les colonnes à droite de cette première colonne sont utilisées pour permettre l'analyse des phénomènes. Chacune correspond à la description de l'amorce de mot en fonction d'un paramètre *(cf.* tabl. 1). Par exemple, le premier paramètre qui a été étudié dès le début des travaux sur ce type de phénomène (Pallaud, 2002) concerne le type d'interruption lexicale. En effet, il a été possible de distinguer trois sortes d'interruptions lexicales selon que le fragment de mot se trouve complété, modifié ou laissé inachevé :

- Fragment complété :
  Tropr102, 3,3 on va **com-** on va **commencer** à les + à le houspiller un peu
- Fragment modifié :
  Tropr101, 1, 3 euh et c'est vrai en plus on (n') a **jam-** on (n') a **pas** compris
- Fragment inachevé :
  Pariscen, 1, 5 alors il y a les O.N.G. + et puis les les **gouver-** les qui qui qui donnent des livres gratuitement

**Tableau 1**
*Exemple d'extrait du tableau obtenu sur* EXCEL

| n° | Corpus | Page, ligne | Fragment de mot et son contexte | Type de troncation | Catégorie grammati-cale | Type de mot | Rupture syntaxique |
|---|---|---|---|---|---|---|---|
| 298 | Pariscen | 2,2 | donc j'ai j'ai proposé un plan de formation + *euh* **se-** *euh* **segmenté** + par euh avec tous les différents savoir-faire de l'éditeur | complétée | verbe | mot plein | continuité |
| 276 | Tropr102 | 3,7 | l'Italie a d'ailleurs a fait une loi - pour les repentis -- qui est bon qui était très contestée etc. mais qui a porté **beauc-** qui a **bien** porté ses fruits | modifiée | adverbe | mot-outil | continuité |
| 274 | Tropr102 | 3,5 | il y avait un reportage à la télévision là que j'avais vu ça à la **tél-** c'était vachement intéressant | inachevée | nom | mot plein | rupture syntaxique |

Pour chaque énoncé, plusieurs paramètres peuvent être distingués et leurs interactions étudiées. Par exemple, pour l'étude des effets dans le contexte droit par la troncation de mot (Pallaud, 2006), les paramètres observés ont été les suivants :

- la position syntaxique du fragment et sa localisation dans le syntagme
- la localisation de la reprise, quand il y en a une, après l'interruption lexicale
- la poursuite ou l'interruption de l'énoncé après la reprise

Le tableur EXCEL permet des quantifications et des représentations graphiques rapides. Un des objectifs de cette étude était de préciser l'influence du type de troncation (amorce complétée, modifiée et inachevée) sur la poursuite ou l'interruption de l'énoncé après la reprise. La fonction-

filtre de ce tableur permet de croiser ces deux variables et de déterminer par un Khi² [14], par exemple, quels sont les effets de l'une sur l'autre. A quelques exceptions près, les trois-quarts des amorces de mots complétées ou modifiées ne sont pas suivies d'une rupture syntaxique. Il n'en est pas de même pour le quart restant, laissé inachevé : 65% de ces troncations sont suivies d'une rupture de construction verbale (fig. 1).

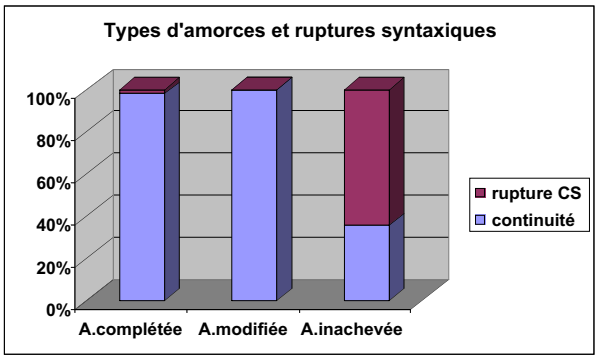


**Figure 1**
*L'influence du type de troncation sur le devenir de la proposition*

Par ailleurs, l'étude de la position syntaxique de la troncation de mot a consisté à déterminer sa localisation avant (ou sur le verbe) ou après le verbe recteur. Cette variable a également une influence sur l'impossibilité pour le locuteur de maintenir la continuité de son énoncé. Les cas de rupture de construction verbale sont deux fois plus nombreux lorsque la troncation de mot se place avant et sur le verbe qu'après le verbe. Autrement dit, plus la troncation a lieu tôt (avant et sur le verbe), plus l'énoncé risque de s'interrompre[15]. En revanche, les cas de continuité syntaxique sont aussi nombreux, quelle que soit la localisation du fragment de mot.

14. Le test du Khi² permet de décider avec une certaine probabilité si des échantillons (ici fragments de mot complétés, modifiés ou inachevés) appartiennent à une même population au regard du type de mot, par exemple. Les résultats de cette analyse sont négatifs :un mot plein a autant de chance d'être complété qu'un mot-outil. En revanche, le mot plein sera plus souvent modifié et moins souvent laissé inachevé que le mot-outil (Pallaud, 2004). Ces deux types de mot ne sont donc pas, statistiquement, de la même population.

15. Khi²=4,40 ; $p<.05$ ; d.d.l.=1 ; le résultat du test montre que la différence est significative avec un risque d'erreur de 5%.

## 5. Conclusion et projet

Le commentaire qui accompagne la production d'un *lapsus* souligne la surprise du locuteur : « ça m'a échappé ». Or, ce que montrent les études actuelles sur les corpus oraux est que la parole semble bien échapper le plus souvent au locuteur et que cela n'a rien de pathologique. L'énoncé ne peut être produit que de cette façon : le locuteur égrène en le découvrant son énoncé telle l'araignée qui sécrète au fur et à mesure le fil auquel elle est suspendue. Parler témoigne du maillage des mots qui surgissent à l'appel de ce qui est déjà dit et va être dit. Le locuteur poursuit son énoncé tout autant en le complétant qu'en le rattrapant mot après mot et même en l'anticipant. L'équivocité du terme *poursuivre,* souvent employé à propos de l'énonciation (« poursuis ce que tu as commencé à dire ») convient bien à cette situation dont la caractéristique première est d'être équivoque.

Le projet à court terme est d'augmenter la base de données des troncations involontaires de mots grâce aux vingt autres corpus disponibles et surtout de lier cette base au contexte de son corpus et à la source sonore. Ce montage permettrait de pouvoir, à tout moment, réécouter l'énoncé choisi (l'extrait contenant le fragment de mot) sans avoir à revenir au corpus initial. Cette base de données contiendrait donc les extraits transcrits comportant une amorce de mot et la version sonore de ces extraits.

## 6. Références bibliographiques


ARNAUD, P.J.L., (1997), Les ratés de la dénomination. Typologies dans les lapsus, in Boysson & Thoinon (éds), *La dénomination,* PUL, Lyon, 307 p.

BLANCHE-BENVENISTE, C. ; JEANJEAN, C. (1987). *Le français parlé. Transcription et édition,* Paris : Didier Erudition.

CLARK, H. ; WASOW, T. (1998). Repeating words in spontaneous speech, *Cognitive Psychology,* 37*,* p. 201-242.

DUEZ, D. (2001). Signification des hésitations dans la production et la perception de la parole spontanée, *Revue Parole,* 17-18-19, p. 113-138.

FENOGLIO, I., (1997). La notion d'évènement d'énonciation : le « lapsus » comme une donnée d'articulation entre discours et parole, *Langage et société,* 80, p. 39-71.

GADET, F. (2000). Derrière les problèmes méthodologiques du recueil des données, *Les Cahiers de l'Université de Perpignan*, 31, « Linguistique sur corpus », p. 30-43.

GADET, F. (2003). Derrière les problèmes méthodologiques du recueil des données, *Texto !* juin-septembre, <http://www.revue-texto.net/Inedits/Gadet_Principes.html>, consultée le 18/07/2006).

HENRY, S. (2002). Étude des répétitions en français parlé spontané pour les technologies de la parole, *RECITAL'02*, Nancy, France, p. 467-476.

MERINGER, R. & MAYER, K. (1895). *Versprechen und Verlesen: eine Psychologish-linguistische Studie*. Stuttgart : Göschen.

MERINGER, R. (1908). *Aus dem Leben der Sprache : Versprechen, Kindersprache, Nachahmungstrieb,* Berlin : Behr's Verlag.

MERINGER, R. (1923). Die täglichen Fehler im Sprechen, Lesen und Handeln, *Wörter und Sachen,* 8, p. 122-140.

PALLAUD, B. (1999). Lapsus et phénomènes voisins dans la langue parlée : problèmes d'identification, *Recherches sur le Français Parlé,* 15, p. 1-33.

PALLAUD, B. (2001). Les lapsus : des pierres dans le champ linguistique, in M. Arrivé et C. Normand (éds), *Linguistique et Psychanalyse,* Colloque de Cerisy-la-Salle, 1-8 septembre 1998, IN Presse, p. 47-66.

PALLAUD, B. (2002). Erreurs d'écoute dans la transcription de données orales. Actes du colloque *Transcription de la parole normale et pathologique*, Tours, 8-9 décembre 2000, *Revue Parole,* 22-23-24*,* p. 267-294.

PALLAUD, B. & HENRY, S. (2004). Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé, in *Le poids des mots, Actes des 7es Journées Internationales d'Analyse statistique des Données Textuelles,* Louvain-la-Neuve, 10-12 mars 2004, Louvain : PUL, vol. 2, p. 848-858.

PALLAUD, B. (2006). Troncations de mots, reprises et interruption syntaxique en français parlé spontané, in *JADT 2006, 8es Journées internationales d'Analyse statistique des Données Textuelles,* 20-22 avril, Besançon, p. 707-715.

PASDELOUP, V. (1992). A Prosodic Model for French Text-to-Speech Synthesis: A Psycholinguistic Approach, in Bailly, G. ; Benoit, Chr. ; Sawallis, Th.R. (eds), *Talking Machines. Theories, Models and Designs,* p. 335-348.

PILLON, A. (1998). Morpheme units in speech production evidence from laboratory-induced verbal slips, *Language and cognitive processes,* 13, 4, p. 465-498.

PETER-DEFARE, E. (1993). *Aspects phonologiques, syntaxiques et phonologiques de l'empan des erreurs de langage,* DEA de Phonétique, Université de Provence, Aix-en-Provence, 166 p.

ROSSI, M. & PETER-DEFARE, E. (1995). *Lapsus linguae:* word errors or phonological errors?, *International Journal of Psycholinguistics,* 11, 1[30]*,* p. 5-38.

ROSSI, M. & PETER-DEFARE, E. (1998). *Les lapsus ou comment notre fourche a langué,* Paris : PUF.

SHRIBERG, E. (1999). Phonetic Consequences of Speech Disfluency, Symposium on the Phonetics of Spontaneous Speech (S. Greenberg, P. Keating, organizers), *Proc. International Congress of Phonetic Sciences,* San Francisco, 1, p. 619-622.